

Disk controller having host interface and bus switches for selecting buffer and drive busses respectively based on configuration control signals

Patent Number: ☐ US5257391
Publication date: 1993-10-26
Inventor(s): DULAC KEITH B (US); WEBER BRET S
Applicant(s): NCR CO (US)
Requested Patent: ☐ JP5197495
Application US19910746399 19910816
Priority Number(s): US19910746399 19910816
IPC Classification: G06F13/00
EC Classification: G06F3/06D, G06F11/10M
Equivalents: JP3204276B2

Abstract

A disk array controller providing a variable configuration data path between the host system and the individual disk drives within a disk array and parity and error correcting code generation and checking. The controller includes host interface logic for converting data received from the host system via a 16 or 32-bit SCSI bus to 16, 32 or 64-bit data words multiplexed across one, two or four 16-bit buffer busses, and for converting data received from the buffer busses to the proper form for transmission to the host system. A bus switch, including an exclusive-OR circuit for generating parity information, is connected between the buffer busses and six disk drive busses for directing the transfer of data and parity information between selected buffer and drive busses. The controller further includes a storage buffer connected to the buffer busses to provide temporary storage of data and parity information. The host interface logic, bus switch and storage buffer, under the direction of an included processor and DMA control logic, performs array read and write operations requested by the host system in accordance with RAID level 1, 3, 4 or 5 protocols.

Data supplied from theesp@cenetest database - I2

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平5-197495

(43) 公開日 平成5年(1993) 8月6日

(51) Int.Cl.⁵

G 0 6 F 3/06

識別記号

3 0 1 B

庁内整理番号

7165-5B

R 7165-5B

F I

技術表示箇所

審査請求 未請求 請求項の数14(全 22 頁)

(21) 出願番号 特願平4-237616

(22) 出願日 平成4年(1992) 8月14日

(31) 優先権主張番号 7 4 6 3 9 9

(32) 優先日 1991年 8月16日

(33) 優先権主張国 米国 (U S)

(71) 出願人 592089054

エヌ・シー・アール・インターナショナル・インコーポレイテッド

アメリカ合衆国 45479 オハイオ、デイトン サウス バターソン プールバード 1700

(72) 発明者 キース ビー、デュラーク

アメリカ合衆国 67037 カンザス、ダービー、ヒラ 8652

(72) 発明者 ブレット エス、ウェバー

アメリカ合衆国 67203 カンザス、ウィチタ、エヌ、マウント カーマル 1851

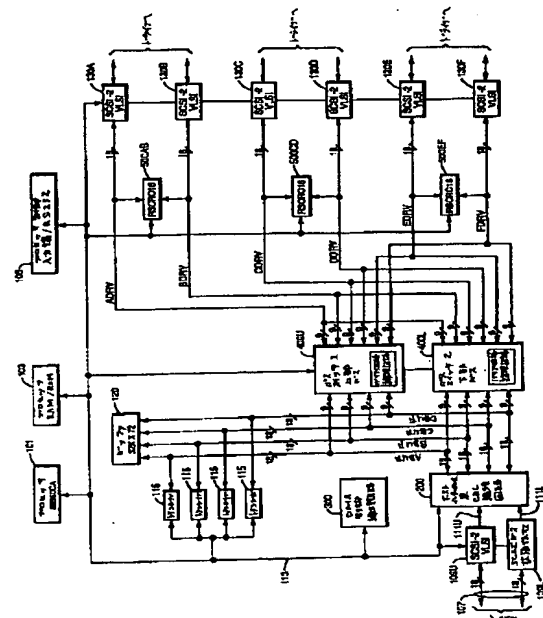
(74) 代理人 弁理士 西山 善章

(54) 【発明の名称】 ディスクアレーコントローラのアーキテクチャ

(57) 【要約】

【目的】 ホストシステムおよびディスクアレー内のディスクドライブ間に可変データ路を構築し、パリティおよびエラー矯正コードの発生と検査を行うディスクアレーコントローラを与える。

【構成】 このコントローラは、SCSIバス111を介してホストシステムから受信したデータを多重データワードに変換すると共に、バッファバスABUF-DBUFから受信したデータをホストシステムへの送信のため適切な形態に変換するホストインターフェース論理回路200を含む。選択したバッファおよびドライブバス間におけるデータおよびパリティ情報の転送を監督するためのパリティ情報発生用排他的OR回路付きバススイッチ400U、400Lが設けられる。ホストインターフェース論理回路、バススイッチ、および格納バッファ120は、プロセッサ101およびDMA制御論理回路300の監督の下にRAIDレベル1、3、4もしくは5プロトコルに従ってホストシステムが要求するアレー読み取り/書き込みオペレーションを行う。



(2)

特開平5-197495

1

2

【特許請求の範囲】

【請求項1】 ホストコンピューターシステムバスを複数のRAIDディスクドライブバスに接続するコントローラであって、

複数のバッファバスと、

該ホストバスおよび該複数のバッファバス間に接続され、該ディスクアレーのRAIDレベル形態に基づいてバッファバスを選択すると共に該ホストバスおよび該選択されたバッファバス間でデータを転送する、ホストインターフェースと、

該複数のバッファバスおよび該複数のドライブバス間に接続され、該ディスクアレーRAIDレベル系他に基づいてバッファバスを選択すると共に、該ホストバスおよび該選択されたバッファバスと該選択されたドライブバスとの間でデータを転送する複数のバススイッチと、を含むコントローラ。

【請求項2】 複数のディスクドライブバスに対しホストコンピューターシステムバスをインターフェースするコントローラであって、

複数のバッファバスと、

該ホストバスおよび該複数のバッファバス間に接続され、該ホストバスと該複数のバッファバスのうちから選択された少なくとも一つのバッファバスとの間でデータを転送するホストインターフェースと、

該複数のバッファバスおよび該複数のドライブバスとの間に接続され、一群の選択されたバッファバスを一群の選択されたドライブバスに結合するバススイッチとを含むコントローラ。

【請求項3】 請求項2に記載のコントローラであって、さらに該複数のバッファバスに接続され、該バッファバスから受信したデータを格納すると共に格納したデータを該バッファバス上に出力するバッファ格納装置を含むコントローラ。

【請求項4】 請求項3に記載のコントローラにおいて、該ホストインターフェースが該ホストバスおよび該バッファバスとの間でデータを指向させる多重化手段を含むことを特徴とするコントローラ。

【請求項5】 請求項3に記載のコントローラにおいて、該バススイッチが、

該複数のドライブバスの各々に対応するバスマルチプレクサからなる第一の複数のバスマルチプレクサ群にして該マルチプレクサ各々がその対応のバスに接続される出力端と複数入力端とを有し、該入力端各々が該複数のバッファバスのなかの対応する一バッファにそれぞれ接続される、前記第一の複数のバスマルチプレクサ群を含むことを特徴とするコントローラ。

【請求項6】 請求項5に記載のコントローラにおいて、該バススイッチがさらに、複数入力端を有するパリティ発生回路にしてこれら入力端の各々が該複数のバッファバスのなかの対応する一バッファに接続されているパリテ

ィ発生回路を有し、

該第一の複数のバスマルチプレクサ各々が、該パリティ発生回路の出力端に接続される入力端を有することを特徴とするコントローラ。

【請求項7】 請求項3に記載のコントローラにおいて、該バススイッチがさらに、

該複数のバッファバスの各々にそれぞれ対応するバスマルチプレクサからなる第二の複数のバスマルチプレクサ群にして、該第二の複数のバスマルチプレクサの各々がその対応のバスに接続される出力端と複数の入力端とを有し、これら入力端が該複数のドライブバスのなかの対応する一バスにそれぞれ接続されている、前記第二の複数のバスマルチプレクサ群を含むことを特徴とするコントローラ。

【請求項8】 請求項7に記載のコントローラにおいて、該バススイッチがさらに、

複数の入力端を有するパリティ発生回路にしてこれら入力端が該複数のドライブバスのなかの対応する一バスに接続されるパリティ発生回路を有し、

該第二の複数のバスマルチプレクサの各々が、該パリティ発生回路の出力端に接続される入力端を有することを特徴とするコントローラ。

【請求項9】 請求項8に記載のコントローラにおいて、

該第二の複数のバスマルチプレクサが、一出力端と複数入力端とを有し、これら入力端が該複数のドライブバスの対応する一バスに接続されており、

該パリティ発生回路がこれらバスマルチプレクサの出力端に接続された入力端を含むことを特徴とするコントローラ。

【請求項10】 請求項3に記載のコントローラであって、該ドライブバスの各々に関連されたCRC論理回路にして該ドライブバス上に転送されたデータに対するエラー矯正コードを発生し検査するCRC論理回路をさらに含むコントローラ。

【請求項11】 請求項10に記載のコントローラにおいて、該CRC論理回路が、

ディスクアレー書き込みオペレーション期間中の第一モードではエラー矯正コードを発生すると共に該ドライブバスにおけるデータ転送に該コードを付記すべく動作し、

ディスクアレー読み取りオペレーション期間中の第二モードではエラー矯正コードを発生すると共にそのコードを該ドライブバスにおけるデータ転送に付随するエラー矯正コードと比較すべく動作することを特徴とするコントローラ。

【請求項12】 請求項3に記載のコントローラであって、

該コントローラがさらに、該ホストインターフェース手段、該バススイッチ手段、該バッファ格納手段、アドレス/データバスおよび複数の制御線と相互接続されたブ

(3)

特開平5-197495

3

4

ロセッサおよびDMA制御論理回路を含み、
該プロセッサが該DMA制御論理回路、該ホストインターフェース手段、該バススイッチ手段、および該バッファ格納手段に制御信号を与え、該DMA制御論理回路、該ホストインターフェース手段、該バススイッチ手段、および該バッファ格納手段の構築および動作が該制御信号により決定され、

該DMA制御論理回路が該ディスクアレーコントローラのためのDMAおよびバッファ制御を与えることを特徴とするコントローラ。

【請求項13】複数のディスクドライブに対しホストコンピュータシステムをインターフェースするディスクアレーコントローラであってホストバスと、

複数のバッファバスと、

該ホストバスおよび該複数のバッファバス間に接続され、該ホストバスと該複数のバッファバスのなかから選択された少なくとも一バッファバスとの間でデータを転送するホストインターフェース手段と、

複数のドライブバスと、

該複数のバッファバスと該複数のドライブバスとの間に接続され、一群の選択されたバッファバスを一群の選択されたドライブバスに結合するバススイッチ手段と、

該複数のバッファバスに接続され、該バッファバスから受信したデータを格納すると共に格納したデータを該バッファバス上に出力するバッファ格納手段と、

排他的OR回路にして、

該バッファおよびドライブバスの中から選択したものからデータを受信すべく該排他的OR回路を接続すると共に該選択したバスから受信したデータを結合する、第一スイッチ手段と、

該バッファおよびドライブバスの中から選択したものに該排他的OR回路の出力を与える第二スイッチ手段とを含む排他的OR回路とを含むディスクアレーコントローラ。

【請求項14】請求項12に記載のコントローラにおいて、

該ホストバスが16ビットSCSIバスを含み、

該複数のバッファバスが四つの16ビットバッファバスを含み、

該複数のドライブバスが六つの16ビットドライブバスを含み、

該コントローラがさらに、六つのドライブバスに対応する六つのSCSIアダプタを含み、該アダプタ各々がその対応するドライブバスおよび対応するディスクドライブ間に接続されていることを特徴とするコントローラ。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明はコンピュータシステム用ディスク格納デバイスに関し、特に複数の格納デバイスを管理するディスクアレーコントローラに関する。

10

20

30

40

50

【0002】

【従来の技術】過去数年の間にコンピュータの一次メモリデバイスの速度および容量は、その処理能力および処理速度と共に劇的な進歩を遂げた。しかし不幸なことにデータ処理と一次メモリ技術の進歩とは対照的に、主に磁気ディスクからなる二次メモリ格納デバイスの性能の改良はそれほど進捗してない。このようなCPUおよび一次メモリデバイスの性能の向上および速度の向上は、それが実現され続けているとしても二次格納デバイスの性能の向上と一致しないならば、浪費されてしまう。例えばCPUと二次格納装置の性能の不一致の結果として、ディスクI/OオペレーションがCPUの多大な比率を占めることとなる。

【0003】ディスクアレーは、二次格納デバイスの性能を改良するための一手段として提案されたもので、その目的はCPUの性能と二次格納装置の性能との間の不一致が高価につくのでこれを解消するためであった。並列接続された複数の小型かつ廉価なディスクドライブを含むディスクアレーは、ホストシステムにとっては単一の大型高速ディスクのように動作する一方で、単一の大型磁気ディスクよりも性能、信頼性、消費電力、および拡張/縮小性において一層改善できる。多くの場合、単一の大型磁気ディスクよりも性能、信頼性、消費電力、および拡張/縮小度においてディスクアレーが優れている。

【0004】現行のディスクアレーの設計で採り得る方法は、デービッド エイ パターソン、ガース ギブソン、およびランディー エッチ カッツ 共著の「廉価なディスク冗長アレーの例 (Redundant Arrays of Inexpensive Disks, RAID)」と題するユニバーシティ・オブ・カリフォルニア レポート 第UCB/CS D87/391号(1987年12月号)に記載されている。この記事はRAIDレベルと呼ばれる五段階のディスクアレー構成を紹介している。最も簡単な構成は、RAIDレベル1システムで、これは一つ以上のデータ格納用ディスクと、当該データディスクに書き込まれた情報のコピーを格納するための同数の追加的「鏡像」ディスクを含む。残りのRAIDレベルはいくつかのデータディスクにまたがってデータを格納する。エラー検査情報およびパリティ情報の格納のためには一つ以上の追加ディスクが利用される。

【0005】タンデムコンピュータインコーポレーテッド社から1990年1月に発行されたタンデムテクニカルレポート第90.2号に記載の「ディスクアレーのパリティストリップング：許容可能なスループットを具えた廉価かつ高信頼性格納装置」と題するジム グレー、ボブ ホースト およびマーク ウォルカー共著の記事には別のディスク構成が記載されている。このパリティストリップングシステムでは、パリティ情報のみがディスク間に分配され、パリティは大きな連続的拡がり

(4)

特開平5-197495

5

をもつようにマッピングされる。データはディスク間に分割されず、在来方法で格納される。

【0006】アレー内の多重ディスクの読み取りおよび書き込みを行うべく多重ディスクドライブを組織化するためには、パリティの発生と検査、データの回復と再構築、および複雑な格納管理技術が必要とされる。これら前記先行技術文献に記載されたディスクアレーシステムの多くは、ホストがRAIDコントローラとして動作し、パリティの発生と検査その他の格納管理動作を行う。これらの機能をホストに行わせることは、ホストのプロセッシングオーバーヘッドとなる点で高価につく。

【0007】さらに先行技術システムの多くは複数ディスクドライブをホストシステムに相互接続している固定データ路構造を含んでいる。いろいろの個数のディスクドライブあるいはいろいろのRAID構造を許容するようにディスクアレーシステムを再構築することは、容易には達成できない。

【0008】

【発明が解決しようとする課題】それゆえ、本発明は新規かつ改良されたディスクアレーコントローラを与えることを課題とする。

【0009】本発明のもう一つの課題は、ホストシステムとディスクアレーに含まれる個別ディスクドライブとの間に可変データ路を含むアレーコントローラを与えることである。

【0010】本発明の別の課題は、パリティおよびエラー矯正コードの計算と検査を行うための論理回路を含むディスクアレーコントローラを与えることである。

【0011】本発明のさらに別の課題は、いろいろの個数のディスクドライブを許容すべく構築できるディスクアレーコントローラのアーキテクチャを与えることである。

【0012】本発明のさらに別の課題は、いろいろのRAID構造を許容すべく構築できる新規かつ有用な、ディスクアレーコントローラのアーキテクチャを与えることである。

【0013】

【課題を解決するための手段】本発明によれば、複数ディスクドライブにホストコンピュータをインターフェース(interface、緩衝)するディスクアレーコントローラが与えられる。このコントローラは当該ホストシステムに接続されたホストバスと、当該ディスクアレー内の個別ディスクドライブに関連した複数ドライブバスとの間のデータ通信を監督する。このコントローラは、ホストバスと、ホストバスおよび選択された一以上のバッファバス間でデータ転送するための複数バッファバスとの間に接続されたホストインターフェース手段を含む。このコントローラはさらに、選択された一群のバッファバスを選択された一群のドライブバスに結合するための

6

ドライブバスと前記バッファバスとの間に接続されたバススイッチを含む。このバススイッチはパリティ発生器を含み、その出力は選択された任意のバッファまたはドライブバスに指向することができる。これらバッファバス上に置かれたデータを格納するとともに格納されたデータをバッファバス上に出力するため、格納バッファが前記複数バッファバスに接続される。

【0014】上記コントローラアーキテクチャは、内蔵プロセッサおよびDMA制御論理回路の監督の下に、ホストシステムがRAIDレベル1、3、または5のプロトコルに従って要求したアレー読み取りおよび書き込みオペレーションを行う。

【0015】本発明の上記その他の課題、特徴、および利点は以下の説明および添付の図面から明かとなろう。

【0016】

【実施例】図1を参照すると本発明の好ましい実施例であるディスクアレーコントローラのアーキテクチャがとってブロック線図で表されている。このコントローラはホストインターフェース兼CRC論理回路ブロック200を介してホストコンピュータシステム(図示して無し)とデータ交換する。ホストインターフェース論理ブロック200はプロセッサ101の制御の下に、このホストシステムに関連する18ビットもしくは36ビット幅の外部SCSI-2バス107とこのホストシステムの四つの18ビット幅内部バッファバスABUF、BBUF、CBUF、DBUFとの間をインターフェースする。バス107は、ブロック109U、109Lで表す標準的SCSIチップのセット(組)および18ビットバス111U、111Lを介してホストインターフェース兼CRC論理回路200に接続する(CRCはcyclic Redundancy check サイクリック冗長性検査の略)。ブロック200とプロセッサ101との間の相互接続はアドレス/データバス113で与えられる。ブロック200の内部構造および動作は図2および図3を参照して後で詳述する。

【0017】内部バッファバスABUF、BBUF、CBUFおよびDBUFはホストインターフェース兼CRC論理回路ブロック200をRAMバッファ120並びに上部および下部バイトバススイッチ400U、400Lに接続する。バッファ120は、四つのバッファバスから来る72ビット幅のワード(語)あるいはこれらのバッファバスのうちの任意の一つから来る各18ビット幅ワードの読み取りおよび書き込みを行う能力を有する。バス113への18ビットもしくは36ビットのアクセスもトランシーバ115を介して与えられる。

【0018】バススイッチ400U、400LはバッファバスABUF、BBUF、CBUFおよびDBUFと六つの18ビット幅ドライブバスADRV、BDRV、CDRV、DDRV、EDRV、FDRVと間の可変バスマッピングを与える。これらの各スイッチは一バイト

(5)

特開平5-197495

7

(8ビットデータおよび1ビットパリティ)の情報を経路化(routing)する。バススイッチ400U、400Lはさらに、パリティ情報、検査パリティ情報を発生すると共に故障したディスクドライブに格納された情報を再構築する能力を有する。このパリティ情報は任意のバッファもしくはドライブバス上に指向できる。図6ないし図10は後述するようにバススイッチ400Uおよび400Lの内部構造および動作に関する詳細を示す。

【0019】ドライブバスADRV、BDRV、CDRV、DDRV、EDRV、FDRVの各々は、関連のSCSI-2デバイス130Aないし130Fに接続される。これらのSCSIデバイスはディスクアレーを形成する六つの対応ディスクドライブ(図示して無し)への接続を与える。これらの六つのドライブは文字AないしFで識別することにする。バスADRVとBDRVの間、CDRVとDDRVの間、およびEDRVとFDRVの間には、それぞれ16ビットのリード-ソロモンサイクリック冗長性検査(Reed-Solomon Cyclic Redundancy Check, RSCRC16)論理ブロック500AB、500CDおよび500EFが接続され、本アレーコントローラに供するためのリード-ソロモンCRCのエラー発生/検出信号を与える。

【0020】ホストインターフェース論理回路ブロック200、バススイッチ400U、400L、RSCRC16論理ブロック500AB、500CD、500EF、SCSIデバイス109U、109L、130Aないし130Fの制御はマイクロプロセッサ101が行う。プロセッサ101、関連のプロセッサメモリ103、プロセッサ制御入力端105と上記素子との間の通信はアドレス/データバス113により与えられる。また図に示すようにバス113にはDMA制御論理ブロック300が接続されている。図4、図5に詳細に示すようにこのブロック300内の論理回路はホストインターフェース論理回路ブロック200、バススイッチ400U、400L、SCSIデバイス130Aないし130F、およびプロセッサ101に対するDMA制御を与える。

【0021】ホストインターフェース論理回路ブロック200、DMA制御論理ブロック300、バススイッチ400U、400Lおよび500AB、500CD、500EFに含まれる論理回路およびその動作を以下に説明する。

【0022】図2、図3は一体となって、図1のホストインターフェース論理回路ブロック200に含まれる論理回路をブロック線図で示す。このホストインターフェース論理回路ブロック200は六つの主要な型式の内部論理ブロックである制御兼ステータスレジスタ201、SCSI-2 DMAバスハンドシェーキング(handshaking)論理回路203、バッファバスDMAハンドシェーキング論理回路205、FIFOブロック207Aない

8

し207D、パリティ検査ブロック209U、209L、およびパリティ検査兼CRC発生ブロック211Aないし211Dを含む。

【0023】制御兼ステータスレジスタブロック201は、ホストインターフェース論理回路ブロック200の構築、制御、およびリセットを行うための16ビット制御レジスタをいくつか含む。ブロック201はまたホストインターフェース論理回路ブロック200のステータスを決定するため、コントローラマイクロプロセッサ101によって使用されるステータスレジスタを含む。ブロック201内の制御兼ステータスレジスタへのマイクロプロセッサのアクセスは、多重アドレス/データバスAD(0-7)、データバスD(8-15)、並びにチップ選択信号CS/、読み取り信号RD/、書き込み信号WR/、アドレスラッチイネーブル信号ALE、および中断信号INT/を送信するためのいろいろの制御線により与えられる。バスAD(0-7)およびD(8-15)は図1のアドレス/データバス113に含まれる。

【0024】ブロック203はホストインターフェース論理回路ブロック200とSCSI-2デバイス109U、109Lとの間のDMAを実行するのに必要なSCSI-2 DMAバスハンドシェーキング論理回路を含む。このハンドシェーキング論理回路もまた、SCSIデバイス109U、109LとFIFO207Aないし207Dとの間のデータ多重化および多重化解除を制御する。ブロック203はまた、FIFOが空ステータスかもしくはフルステータスかに応じて、リクエスト/肯定応答(acknowledge)ハンドシェーキングを要調する。バッファバスDMAハンドシェーキング論理回路205も、ホストインターフェース論理回路ブロックと外部バッファバスコントローラとの間のDMA転送を制御するため、同様のハンドシェーキング論理回路を含む。

【0025】四つのFIFOブロック207Aないし207Dは、ホストインターフェース論理回路ブロックとバッファABUF、BBUF、CBUF、およびDBUFとの間のハンドシェーキング依存性を除去するのに利用される。FIFOブロック207A、207Bは各々バス111UとバッファバスABUF、BBUFとの間に接続される。FIFOブロック207C、207Dは各々バス111UとバッファバスCBUF、DBUFとの間に接続される。もしも本コントローラアーキテクチャがSCSIバス拡張デバイス109Lおよび関連のバス111Lを含むなら、ブロック207B、207Dもまたバス111Lに接続される。これらFIFOブロックの構成および動作はブロック201内のレジスタにより制御される。各FIFOブロックは四つの18ビットワード(16ビットのデータおよび2ビットのパリティ)まで格納することができる。

【0026】ブロック209U、209Lはホストインターフェース論理回路ブロックとSCSI-2デバイス10

(6)

特開平5-197495

9

9U、109Lとの間で送信されるすべての情報に対しパリティ検査を与える。これらブロックはデータ転送についてのパリティ情報を発生し、発生したパリティ情報を、当該データと共に送信されるパリティ情報と比較する。

【0027】ブロック211Aないし211Dはホストインターフェース論理ブロックと対応のバッファバスとの間のデータ転送に対するパリティ検査を与える。ブロック211Aないし211Dはまた、DMAデータブロックにCRCデータを発生し付加すると共に、DMAデータブロックのCRCデータを検査し付加除去するように機能する。

【0028】作用上、ホストインターフェース論理回路ブロック200はSCSI-2デバイス109U、109Lと四つのバッファABUF、BBUF、CBUF、およびDBUFとの間のデータを多重化するのに使用される。ブロック200はバス111U、111Lと、次の(1)ないし(3)のバスとの間の多重機能を与える：(1)4+1個のRAIDレベル3のアプリケーションに供する四つのすべてのバッファバス（これらは、回転シーケンス順序 (rotating sequential order) に従って四つのバッファABUF、BBUF、CBUF、およびDBUFにまたがるデータをワードストリッピング (word stripping) することにより行う）、(2)2+1個のRAIDレベル3のアプリケーションに供する、二対のバッファバスの一対（これは、回転シーケンス順序に従って前記対にまたがるデータをワードストリッピングすることにより行う）、(3)RAIDレベル1のアプリケーションおよび単一バスRAID5アプリケーションに供するバスの任意の一つ。

【0029】線図4および5は一体として図1のDMA制御論理ブロック300内に含まれる論理回路を示す。DMA制御論理ブロック300は、四つの主要な部分であるマイクロプロセッサインターフェース301、DMAインターフェース321、バッファインターフェース341、およびCRC制御インターフェース361に分割される。

【0030】マイクロプロセッサインターフェース301は次に掲げる機能を果たすべく設計されたいろいろのインターフェース回路を含む。イ、内部レジスタ読み取りおよび書き取り制御（ブロック303）、ロ、アドレスラッチングおよび復号（ブロック305）、ハ、マイクロプロセッサデータバス制御、ニ、中断発生および制御（ブロック307）、ホ、バッファアクセスに対する待機状態発生（ブロック309）。ブロック301へのマイクロプロセッサのアクセスは、多重化されたアドレス/データバスAD（0-15）、アドレスバスADDR（16-21）、並びに送信アドレスラッチイネーブル信号ALE/、チップ選択信号CS/、読み取り信号RD/、書き込み信号WR/、バッファイネーブル信号

10

BE1/、BE2/、バッファチップ選択信号PBFC S/、バッファ指向信号BDIR、準備完了信号RDY/、および中断信号INT/に対するいろいろの制御線により、与えられる。バスAD（0-15）およびADDR（16-21）は図1のアドレス/データバス113内に含まれる。

【0031】DMAインターフェース321は、ホストインターフェース論理回路ブロック200の緩衝に必要な回路およびバススイッチ400U、400Lを介してSCSI-2デバイス130Aないし130Fを駆動するに必要な回路をすべて含む。このDMAインターフェースは次の機能を行う：イ、アクティブDMAチャンネル間のアービトレーション (arbitration、調停)（ブロック323）、ロ、DMAサイクル信号の発生、ハ、DMAリンクの実行。ブロック321との通信は、ホストDMAリクエスト信号HDREQ、ホストDMAスロープ信号HDSTB/、ターゲットDMAリクエスト信号TDREQ、ターゲットDMAスロープ信号TDS TB、バススイッチ出力イネーブル信号BPOE/、DPOE/、プロセッサラッチイネーブル信号PLE/、およびプロセッサポート出力イネーブル信号PPOE/を送信するための制御線を通して与えられる。

【0032】バッファインターフェース341は、ブロック300をRAMバッファ120にインターフェースする回路を含む。インターフェース341は次の機能を支持する。イ、最大4メガバイトまでのバッファアドレス空間をアドレス指定すること、ロ、バッファ読み取り/書き込みオペレーションの制御、ハ、各DMAチャンネルに対するバッファチップ選択の制御。バッファインターフェース341との通信は、バッファアドレスバスBADDR（0-18）と、バッファ読み取り/書き込み信号BUFRD/、BUFRW/、バッファチップ選択信号ABUFCS/、BBUFCS/、CBUFCS/、DBUFCS/の送信に対する制御線とにより、与えられる。

【0033】CRC制御インターフェース361は外部CRC発生器および検査器に対する制御を与える。インターフェース361は外部CRCチップをイニシャライズするためのリセット信号を与え、外部CRCの検査および発生をイネーブル化する。インターフェース361との通信はホスト解除 (clear)、検査、およびシフトのための信号HCRCLR/、HCRCHK/、HCRCSHFT/、およびターゲットの解除、検査およびシフトの信号TCRCLR/、TCRCHK/、TCRCSHFT/を送信するための制御線により与えられる。

【0034】DMA制御論理ブロック300はまた、ブロック300にタイミング、リセット、およびテスト機能を与えるシステムおよびテストブロック381を含む。

(7)

特開平5-197495

11

【0035】動作上、このDMA制御論理ブロックはホストインターフェース論理回路ブロック200、バススイッチ400U、400L、SCSI-2デバイス130Aないし130F、プロセッサ101に供するDMAおよびバッファ制御を与える。このDMA制御論理ブロックは周辺機器としてマイクロプロセッサ101と通信し、内部レジスタの読み取りと書き込みにより制御される。この論理回路は次に掲げる型式のデータ転送を支持する。イ、ホスト読み取り（バッファ120のデータがSCSIホストに送られる）、ロ、ホスト書き込み（ホストから受信したデータがバッファ120に書き込まれる）、ハ、ターゲット読み取り（ドライブアレーから読まれたデータがバッファ120に書き込まれる）、ニ、ターゲット書き込み（バッファ120のデータがドライブアレーに書き込まれる）、ホ、直接書き込み（SCSIホストから受信したデータがバッファ無しにドライブアレーに送られる）、ヘ、直接読み取り（ドライブアレーから読み取られたデータがバッファ無しにSCSIホストへ送られる）、ト、プロセッサによる読み取り、チ、プロセッサによる書き込み。

【0036】バススイッチ400内に含まれる論理回路は図6に線図で示してある。図示した構造は単一半導体チップ上に形成される。番号481ないし484で示す四つのホストポートはそれぞれ四つのコントローラABUF、BBUF、CBUF、およびDBUFへの接続を与える。番号491ないし496で示すアレーポートはそれぞれ六つのディスクドライブバスADRV、BDRV、CDRV、DDRV、EDRV、FDRVに接続する。バススイッチ400はABUF、BBUF、CBUF、およびDBUFの任意の一つと、ドライブバスADRV、BDRV、CDRV、DDRVの任意の一つとの間の単方向接続を与える。いくつかのコントローラバスおよび同数のドライブバスとの間の多重接続も許される。さらに、このバススイッチは、二つ以上のドライブバスに至る任意コントローラバスの単方向接続を与えることができる。バス453を介して得られるパリティ情報は、ドライブバスの任意の一つに出力することもできる。

【0037】バススイッチ400のアーキテクチャは三つの主要ブロック、すなわちラッチモジュール450、スイッチモジュール460、パリティモジュール470から構成される。ラッチモジュール450、スイッチモジュール460、およびパリティモジュール470の内部構造はそれぞれ図7ないし図10に明らかにされている。図7を見ると、ラッチモジュール450は番号401ないし404を付した四つのラッチを含むことが解かる。これらのラッチはそれぞれバスBPAIN、BPBIN、BPCIN、BPDINからデータを受信し、ラッチしたデータをバスBPAINL、BPBINL、BPCINL、BPDINLを介してスイッチモジ

12

ジュール460に与えるべく接続される。

【0038】ラッチモジュール450はさらに、バスBPAOUT、BPBOUT、BPCOUT、BPDOUT、PARINを介して、スイッチモジュール460からデータを受信すべく接続された五つのバスラッチ411ないし415を含む。ラッチ411ないし414の出力はそれぞれバスBPAOUTL、BPBOUTL、BPCOUTL、BPDOUTLに与えられる。ラッチ415の出力はバスPARINLを介してパリティモジュール460に接続される。

【0039】ラッチ401ないし404および411ないし415は受信したデータをラッチし、もしくは通過させるべくコントローラにより発生された制御信号に応答する。また図7には番号421ないし424および431ないし435で示す、各ラッチの出力端に接続されたパリティ検査回路が示されている。各パリティ検査回路はパリティエラーが検出される度にエラー信号を発生する。

【0040】図8および図9は図6に示すスイッチモジュール460の内部構造を示すブロック線図である。モジュール460は六つの5:1マルチプレクサ441ないし446を含む。各マルチプレクサの対応入力端はバスBPAINL、BPBINL、BPCINL、BPDINL、およびパリティモジュール470の出力端PAROUTに接続される。マルチプレクサの出力端441ないし446はそれぞれバスDPAOUT、DPBOUT、DPCOUT、DPDOUT、DPEOUT、およびDPFOUTに接続される。

【0041】スイッチモジュール460はさらに、番号451ないし454で示す四つの7:1バスマルチプレクサを含む。各マルチプレクサ451ないし454の対応入力端は、バスDPAIN、DPBIN、DPCIN、DPDIN、DPEIN、DPFIN、およびPAROUTに接続される。バスDPAIN、DPBIN、DPCIN、DPDIN、DPEIN、DPFINには6:1マルチプレクサ455の入力端も接続される。マルチプレクサ451ないし455の出力端はそれぞれ、バスBPAOUT、BPBOUT、BPCOUT、BPDOUT、およびPARINに接続される。

【0042】各マルチプレクサ441ないし446は、バスBPAINL、BPBINL、BPCINL、BPDINL、およびPAROUTの任意の一つをこれらマルチプレクサの対応出力バスに接続するため、コントローラにより発生された選択信号に応答する。同様に、マルチプレクサ451ないし455は各々、コントローラにより発生された選択信号に応答して、バスDPAIN、DPBIN、DPCIN、DPDIN、DPEIN、DPFINおよびPAROUTの任意の一つを、これらマルチプレクサの出力端に接続されたバスに接続

(8)

特開平5-197495

13

【0043】パリティモジュール470の内部構造は図10のブロック線図に例示されている。モジュール470は四つの4:1マルチプレクサ461ないし464を含む。マルチプレクサ461はバスBPA INL、BPA OUTL、PAR INL、およびBPA OUTからデータを受信すべく接続される。マルチプレクサ462はバスBPB INL、BPB OUTL、PAR INL、およびBPB OUTからデータを受信すべく接続される。マルチプレクサ463はバスBPC INL、BPC OUTL、PAR INL、およびBPC OUTからデータを受信すべく接続される。マルチプレクサ464はバスBPD INL、BPD OUTL、PAR INL、およびBPD OUTからデータを受信すべく接続される。

【0044】パリティ情報を計算し、検査し、選択したバスについて排他ORビット演算を行うことによりドライブデータが再構築される。モジュール470はマルチプレクサ461、462の出力を結合するための第一排他的OR回路471、マルチプレクサ463、464の出力を結合すべく接続された第二排他的OR回路472、および排他的OR回路471、472の出力を結合するための第三排他的OR回路475を含む。パリティモジュール470の出力は、前記三つの排他的OR回路の出力を受信すべく接続された3:1マルチプレクサ479により与えられる。マルチプレクサ479の出力はバススイッチ460に与えられ、上述したように次いでスイッチ460がこのパリティデータを任意のコントローラに指向させ、もしくはバスを駆動することができる。

【0045】マルチプレクサ461ないし464はコントローラにより発生された選択信号に応答して選択されたデータバスを排他的OR回路471、472に結合する。マルチプレクサ479はコントローラにより発生された選択信号に応答して排他的ORオペレーションに含まれるバスの数を制限する。例えばRAIDレベル3、4、もしくは5の書き込みオペレーションでは四つのバスから受信するデータを結合してパリティを発生できる一方、RAIDレベル1に従って構成されたアレー内に保存されたデータを検査するためにはただ二つのバスを結合すれば足りる。

【0046】図11は図1に示すリード-ソロモン サイクリック冗長性検査(RSCRC16)ブロック500 ABに含まれる論理回路のブロック線図である。RSCRC16ブロックRSCRC16 500 ABはドライブバスADRV、BDRV上のデータ転送のためのエラー検査を支持(support)する。このブロックは三つの主要部分、すなわち、マイクロプロセッサインターフェース501、ADRVバスRSCRC16検査器/発生器503、およびBDRVバス用RSCRC16検査器/発生器505、を含む。

【0047】マイクロプロセッサインターフェース50

14

1は、内部レジスタ読み取りと書き込みの制御、アドレスラッチングと復号、および中断発生と制御という機能を果たすべく設計されたいろいろのインターフェース回路を含む。ブロック501へのマイクロプロセッサのアクセスは、多重化されたアドレス/データバス113、およびチップ選択信号CS/、中断信号INT/、読み取り信号RD/、書き込み信号WR/等の制御信号を送信するための線により与えられる。

【0048】RSCRC16発生器/検査器モジュール503、505はそれぞれバスADRV、BDRVに対するエラー検査を行う。各モジュールは、その対応のドライブバスから来るデータと、対応のドライブインターフェース491、492から来るストローブロック信号ASTB/、BSTB/と、CRC制御インターフェース361から来る検査信号ACHECK/、BCHECK/と、シフト信号ASHIFT/、BSHIFT/を受信すべく接続される。RSCRC16発生器/検査器モジュールのブロック線図は図12に例示されている。モジュール503および505は同一である。

【0049】RSCRC16発生器/検査器モジュールは、データインラッチ(data-in latch)534からデータを受信すべく接続された排他的OR回路(図面のXOR)521と、アキュムレータラッチ522とを含む。ラッチ534はこのモジュールの対応のドライブバスから受信したデータを収容する。アキュムレータラッチ522にはコントローラマイクロプロセッサにより与えられるシードデータ(seed data)またはアルファマルチプレクサ528から得られるフィードバックデータのいずれかを負荷することができる。排他的OR回路521の出力端はアルファマルチプレクサ528への入力端を形成する。

【0050】図12のRSCRC16発生器/検査器モジュールの動作は次の通りである。(1)シードデータがアドレス/データバス113からシードラッチ525中に負荷される。このシード値は特定のエラー検出コード(EDC)特性を生ずるように選択される。(2)ドライブバス上に現われる入力データがデータインラッチ534中にラッチ留めされ、フィードデータがアキュムレータラッチ522中にラッチ留めされる。(3)アキュムレータラッチ522内のデータがデータインラッチ534内のデータと排他的OR演算(XOR演算)されてその結果がアルファマルチプレクサ528に与えられる。(4)一組の非同期論理回路を含むアルファマルチプレクサ528が到来データに予定の固定数を乗算する。(5)アルファマルチプレクサ528の出力がアキュムレータラッチ522中にラッチ留めされる。

【0051】ステップ2ないし4はドライブバス上のデータ転送が完了するまで反復される。発生モードオペレーションにおいてはアキュムレータラッチ522中にラッチ留めされた最終値が、転送されたデータと共にト

15

ランシーバ532を経由して格納のためドライブバスへ与えられる。検査モードにおいてはエラーが全く検出されなかったことを示す全ゼロ条件 (all zero condition) があるか否かについてアルファマルチプレクサ528の最終出力が検査される。論理回路526がゼロ検出オペレーションを行う。

【0052】図示し、上述したコントローラアーキテクチャはホストシステムとアレキサンダーディスクドライブとの間の汎用性ある接続を与える。本コントローラアーキテクチャ内に含まれるデータスイッチおよびデータ処理 (data manipulation) コンポーネントはRAIDレベル1、3、4、5に基づくデータ格納およびデータ取り出しオペレーションを可能にする。また、エラー回復、故障ドライブ上に格納された情報データの再生、および予備ドライブ上へのデータ再構築を支持する。

【0053】上記のオペレーションのいくつかに関し以下に論ずる。

【0054】RAIDレベル3の、4+1個 (四つのデータディスクおよび一つのパリティディスク) の書き込みオペレーションは次のように行われる。バス107經由でホストシステムから受信したデータは18ビットワード (16ビットのデータと2ビットのパリティ情報) に分割され、ホストインターフェース200によりバッファバスABUF、BBUF、CBUF、およびDBUF上に多重化され、バッファ120中に書き込まれる。このデータはバッファ120から除去されるとき、バススイッチ400L、400Uにより四つのドライブバス、例えばバスADRV、BDRV、CDRV、DDRVへ送られる。これらのバススイッチはまた、これら四つのバッファバスから受信した情報の排他的ORビット演算を行うことによりパリティ情報を計算する。計算されたパリティ情報はバスEDRVを経由してディスクドライブEへ送られる。第六ディスクドライブであるドライブFは予備ドライブとして保全される。

【0055】RAIDレベル3読み取りオペレーションはドライブバスADRV、BDRV、CDRV、DDRVを介して四つのデータドライブ、すなわちドライブAないしD、からデータを読み取ることにより行われる。このデータはバススイッチ400L、400U、バッファバスABUF、BBUF、CBUF、およびDBUFを通してバッファ120に与えられる。別のオペレーションにおいてデータがこのバッファから読み取られ、ホストインターフェース200へ送られる。インターフェース200は受信した64ビットのデータをホストシステムへ送信するため、16ビットまたは32ビット (パリティ情報を除く) の型式に変換する。バススイッチ内でパリティドライブEから得た対応の情報がデータドライブから読み込まれたデータとの間で排他的OR演算 (XOR演算) され、ドライブAないしDから読まれたデータのパリティ検査が行われる。

(9)

特開平5-197495

16

【0056】ディスクドライブの一つの故障のためにアクセス不可能であるデータはこの読み取りオペレーションの期間に再生できる。例えば、もしもドライブCが故障したとすると、データドライブA、BおよびDから読み取られたデータはバススイッチパリティ論理モジュール内で、ドライブEから得られたパリティ情報と結合することができ、その結果ドライブCのデータを再生することができる。再生されたデータはバッファバスCBUF上に配され、ドライブA、BおよびDから得たデータと共にバッファバスABUF、BBUFおよびDBUF上に置かれ、バッファ120に与えられ、究極的には必要な変換とホストシステムへの送信を行うため、ホストインターフェース200に与えられる。この代わりとして、コントローラアーキテクチャは再生データを予備ドライブFに与えるよう構成することもできる。

【0057】RAIDレベル5の書き込みオペレーションは読み取り手順および書き込み手順の両方を含む。データ路は、前の (古い) データおよび前のパリティ情報をターゲットデータおよびパリティディスクドライブから最初に読み取るべく構成されなければならない。前のデータおよびパリティ情報はバススイッチパリティ論理モジュール内でXOR演算され、その結果は選択したバッファバスを介して格納バッファ120へ与えられる。このオペレーションの期間中、ホストシステムから受信した新規のデータもまた、バッファ120へ書き込まれる。これらのデータ路は次いで該新規なデータとXOR演算されてバッファ120に格納された結果をバススイッチへ与えるべく構成される。この新規データはターゲットデータドライブへ送られる。新規パリティはバッファ120から受信されるXOR演算済み結果と新規データとをXOR演算することにより発生され、パリティドライブへ送られる。

【0058】RAIDレベル5の読み取りオペレーションは、リクエストされたデータを含むドライブからデータを読み取りこのデータをバススイッチおよびホストインターフェースを介してバス107上に送ることにより行われる。例えば、もしもリクエストされたデータがドライブB内に格納されていれば、ドライブAからバスBDRV、バススイッチ400U、400L、バッファバスBBUF、およびホストインターフェース200を経由してホストバス107につなげるデータ路を本コントローラアーキテクチャが構築する。もしもリクエストされたデータが、当該データを収容するディスクドライブ (すなわちドライブB) の故障のためにアクセス不可能であれば、そのデータは残りのドライブA、C、DおよびE上の対応ロケーションから、バススイッチパリティ論理モジュール中に当該データおよびパリティ情報を読み取ることにより、読み取り期間中に再生できる。その再生されたデータすなわちパリティモジュールのXOR演算出力は、バッファバスBBF (任意に選択したも

17

の)を経由してバッファ120へ、そしてさらにその後ホストインターフェース200およびバス107へ、与えられる。

【0059】二つの独立なRAIDレベル5の読み取り-修正-書き込みオペレーションが、上記コントローラーアーキテクチャによりほぼ同時的に行うことができる。ただしそれは二つのオペレーションが異なるデータドライブおよびパリティドライブへのアクセスを要求する場合に限る。このデータ路はまず、前のデータおよび前のパリティ情報をターゲットデータドライブおよびパリティディスクドライブから読み取るよう構成されなければならない。前のデータおよびパリティ情報、並びにホストから受信した新規データは、バススイッチパリティ論理モジュール内でXOR演算され、その結果すなわち新規パリティ情報は、選択したバッファバスを経由して格納バッファ120へ与えられる。このオペレーションの期間、ホストシステムから受信した新規データもバッファ120に書き込まれる。第二書き込みの対象となるデータおよびパリティ情報も上記の方法によりバッファ120に書き込まれる。読み取り-修正-書き込みオペレーションを完了するため、バッファ120内に格納されている新規データおよびパリティ情報をバススイッチを通してターゲットデータおよびパリティドライブに与えるよう、データ路が構築される。

【0060】

【効果】以上述べたところから、本発明により新規かつ有用なディスクアレーコントローラーアーキテクチャを与えることができることが了解される。このアーキテクチャは大きさの異なる、あるいは構造の異なるディスクドライブを許容するように即構築することができる。

【0061】また本コントローラは上記のようにストリップングオペレーション、パリティ発生およびパリティ検査のほか、格納管理オペレーション、およびこれらの機能からのシステムプロセッサの解除を行うことができる。

【0062】さらにまたまた本コントローラーアーキテクチャ内に含まれるデータスイッチおよびデータ処理 (data manipulation) コンポーネントは、RAIDレベル1、3、4、5に基づくデータ格納およびデータ取り出しオペレーションを可能にする。また、エラー回復、故

(10)

特開平5-197495

18

障ドライブ上に格納された情報データの再生、および予備ドライブ上へのデータ再構築を支持する。

【0063】尚、本発明を上記実施例について説明したが、前記特許請求の範囲内で種々の変更が可能であることを了解されたい。

【図面の簡単な説明】

【図1】本発明の好ましい実施例に基づくディスクアレーコントローラのアーキテクチャを示すブロック線図である。

10 【図2】図1に示すホストインターフェース論理ブロック200に含まれる論理回路のブロック線図の一部である。

【図3】図1に示すホストインターフェース論理ブロック200に含まれる論理回路のブロック線図の残りの一部である。

【図4】図1に示すDMA制御論理ブロック300内に含まれる論理回路のブロック線図の一部である。

【図5】図1に示すDMA制御論理ブロック300内に含まれる論理回路のブロック線図の残りの一部である。

20 【図6】図1に示すバススイッチブロック400U内に含まれる論理回路のブロック線図である。

【図7】図6に示すラッチモジュールのブロック線図である。

【図8】図6に示すスイッチモジュールの内部構造を示すブロック線図の一部である。

【図9】図6に示すスイッチモジュールの内部構造を示すブロック線図の残りの一部である。

【図10】図6に示すパリティモジュールのブロック線図である。

30 【図11】図1に示すリード-ソロモン サイクリック冗長性検査(RSCRC16)ブロック500AB内に含まれる論理回路のブロック線図である。

【図12】図11に示すRSCRC16 発生器/検査器モジュール503のブロック線図である。

【符号の説明】

107 外部SCSI-2バス

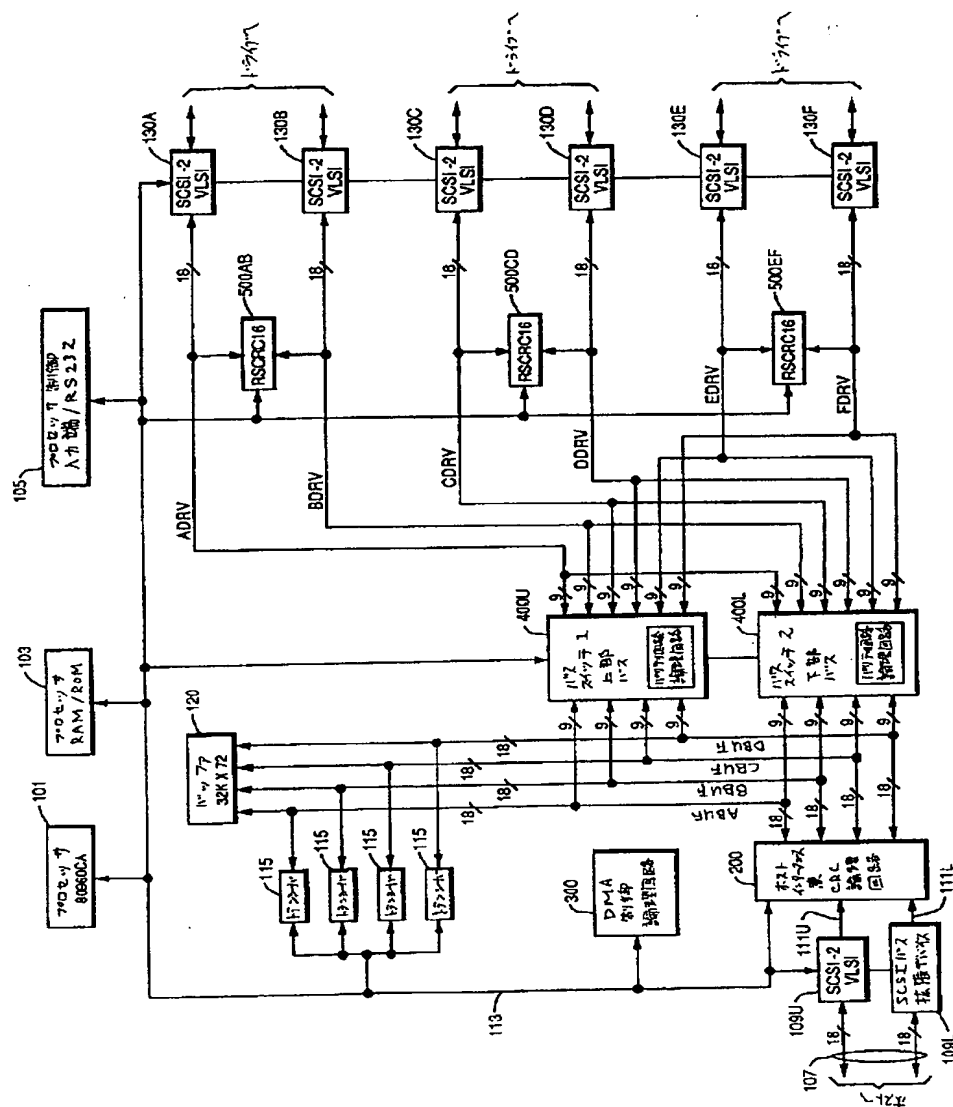
113 アドレス/データバス

500AB、500CD、500EF サイクリック冗長性検査論理ブロック

(11)

特開平5-197495

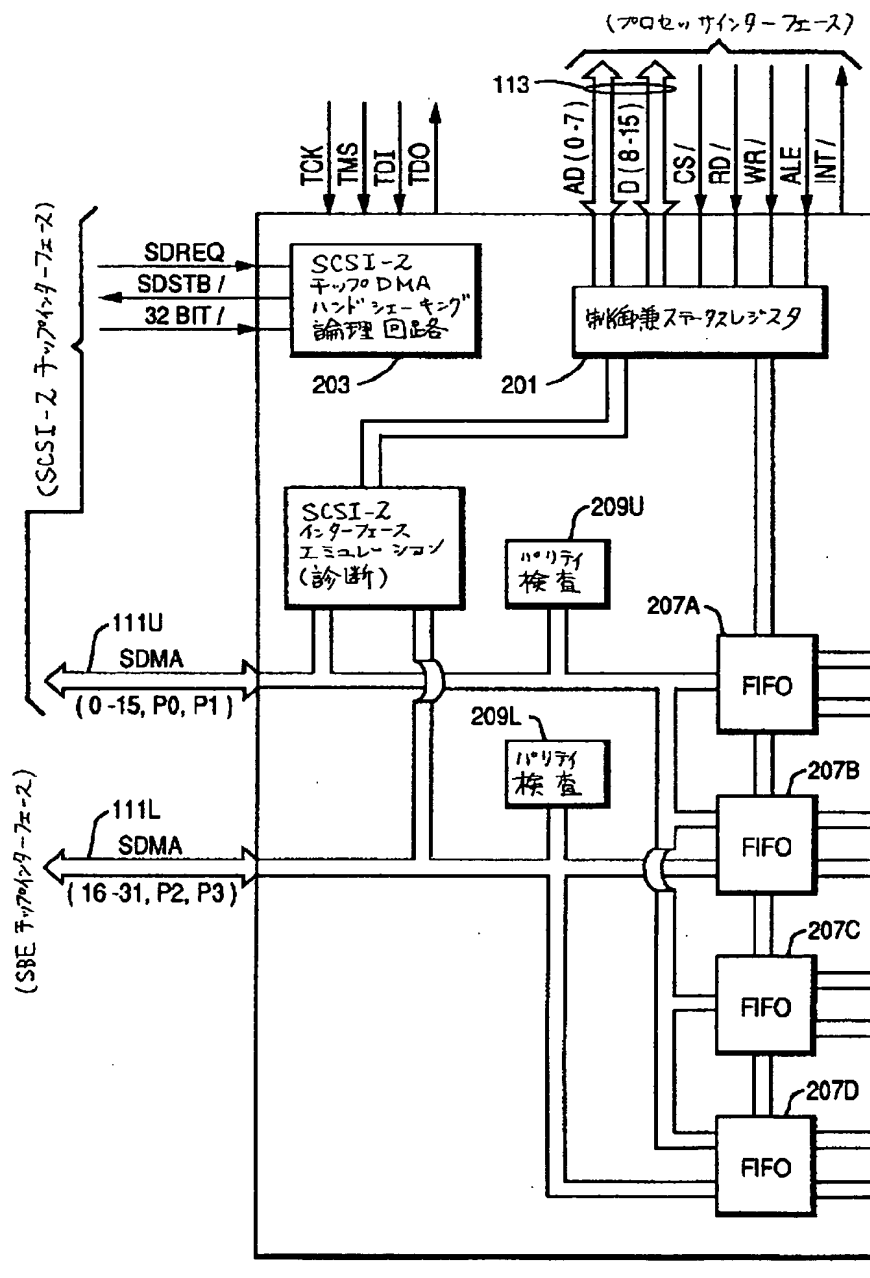
【图 1】



(12)

特開平5-197495

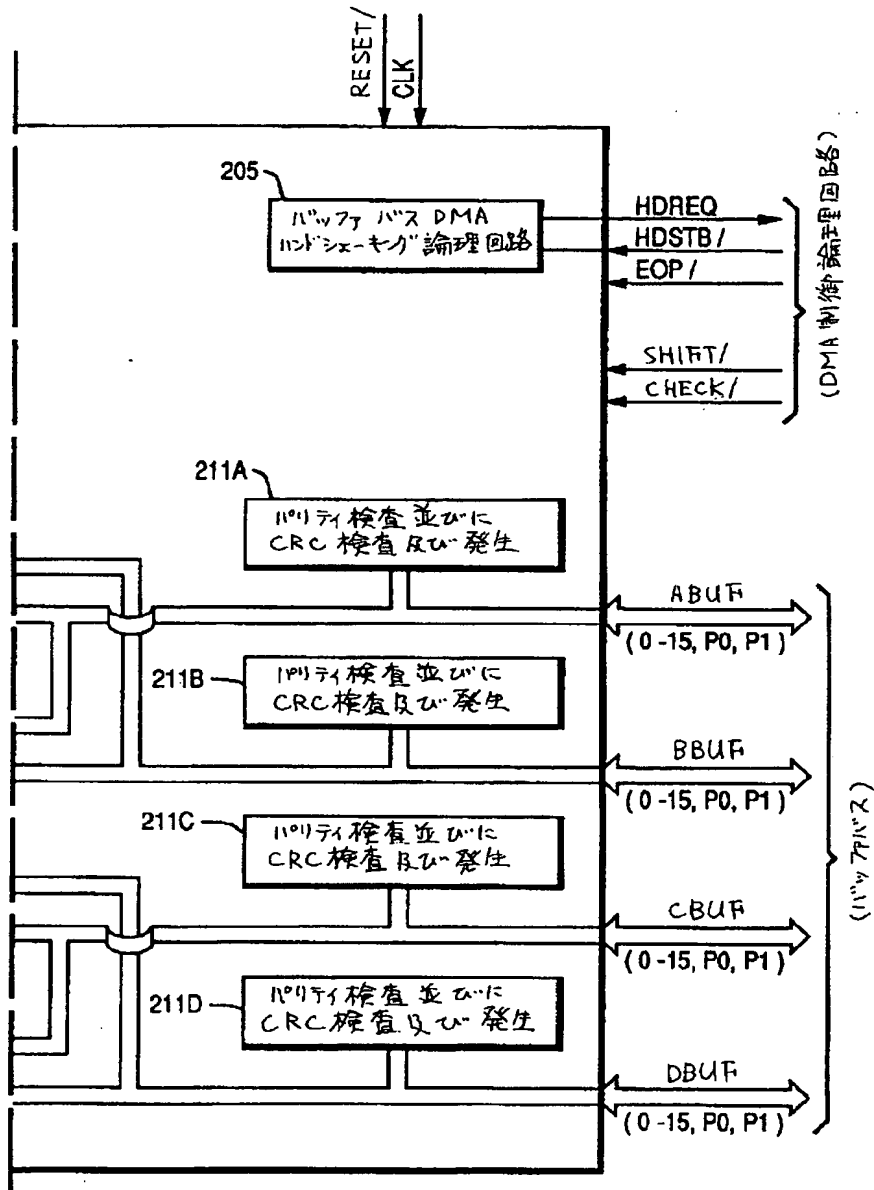
【図2】



(13)

特開平5-197495

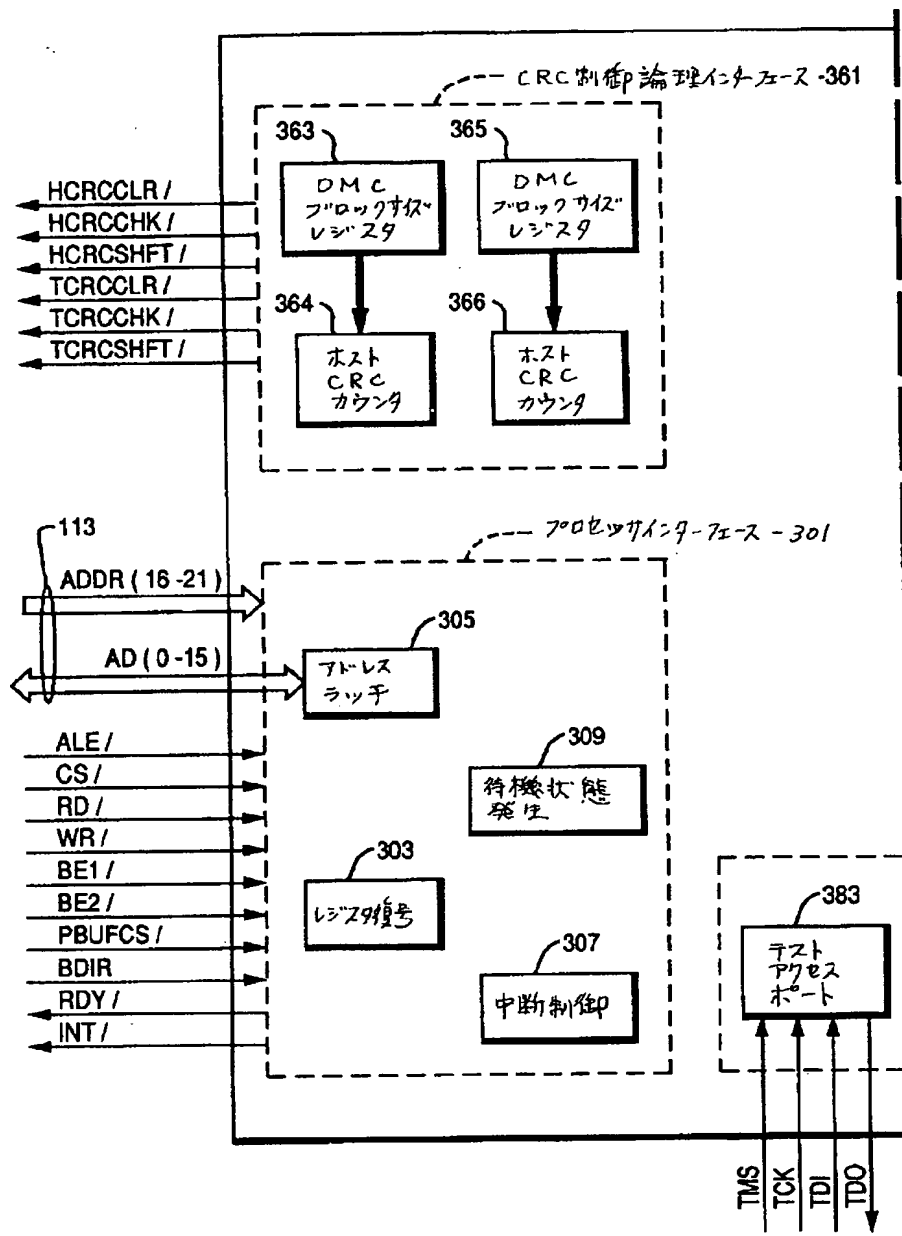
【図3】



(14)

特開平5-197495

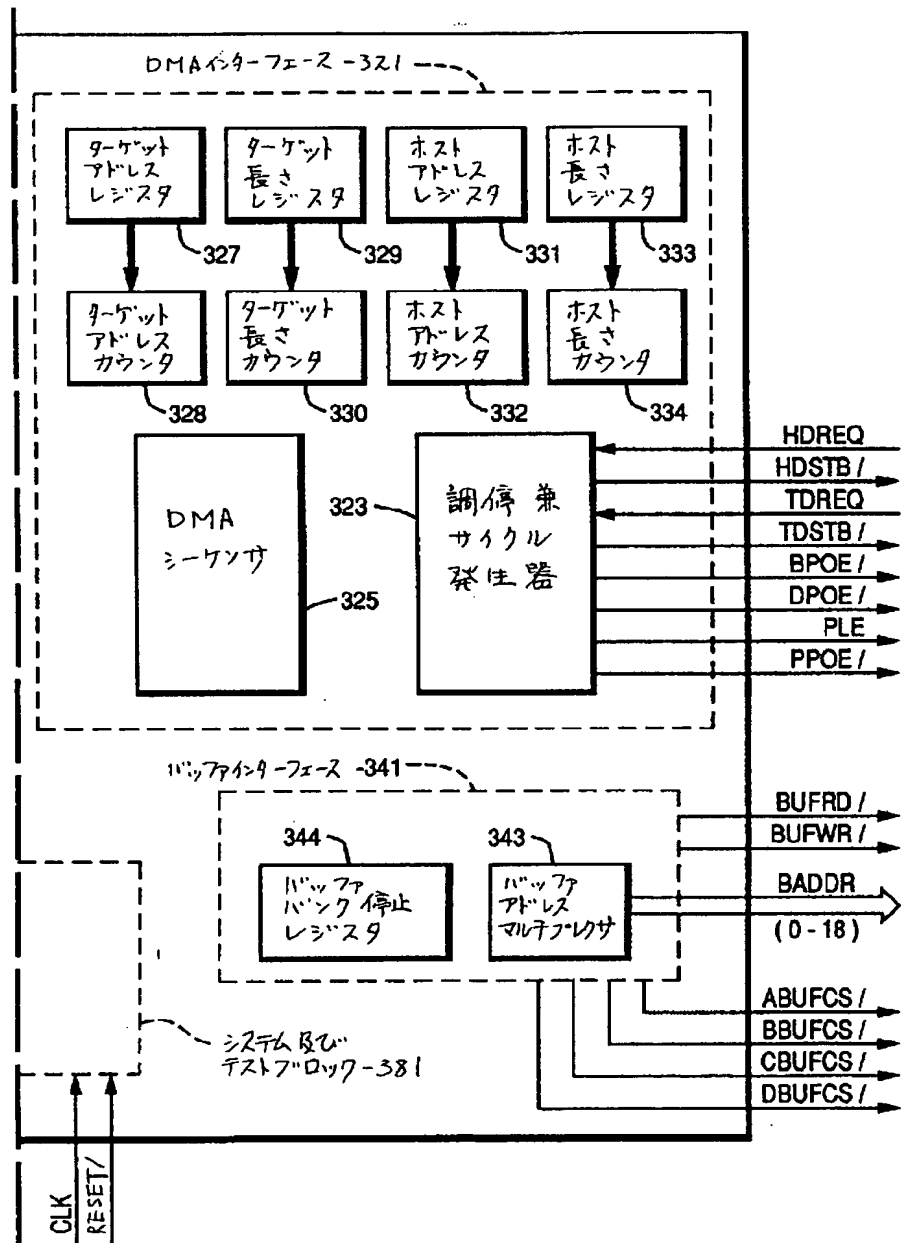
【図4】



(15)

特開平5-197495

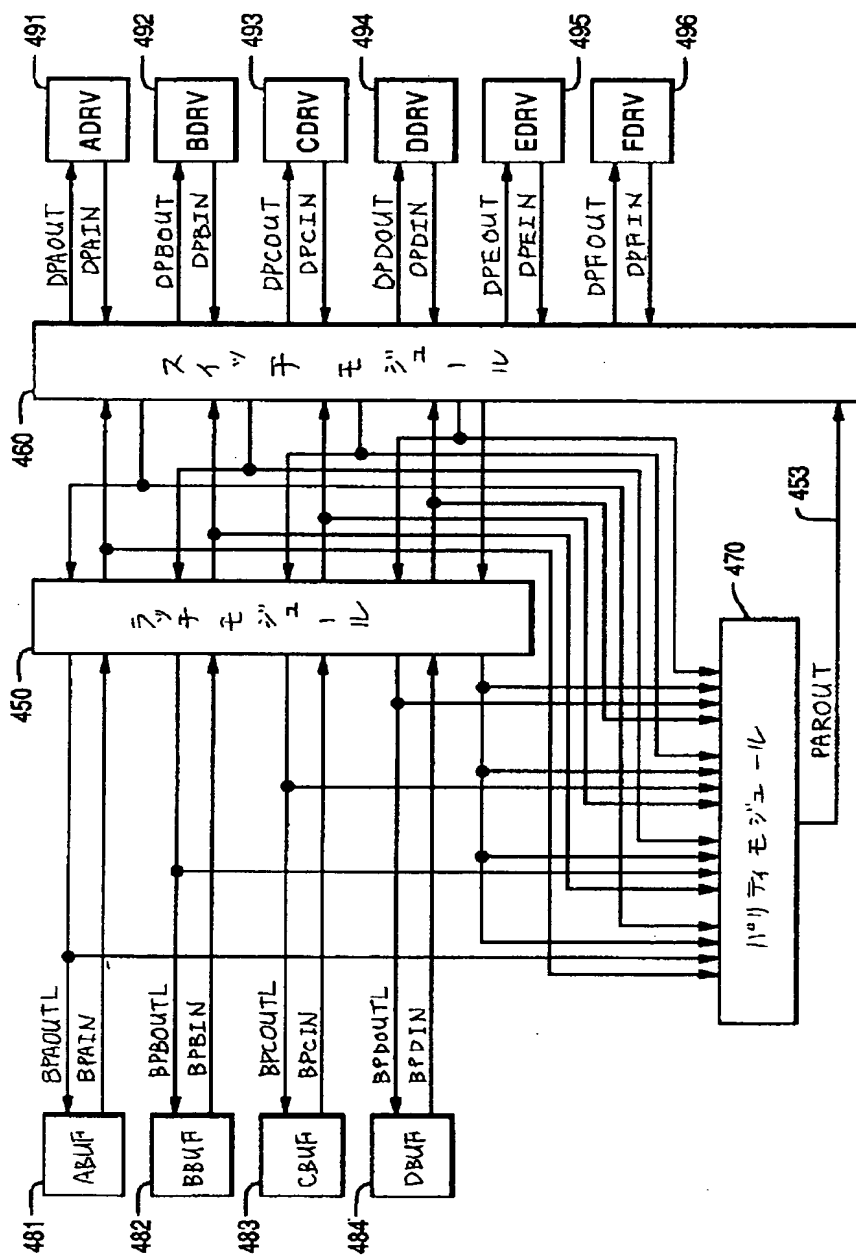
【図5】



(16)

特開平5-197495

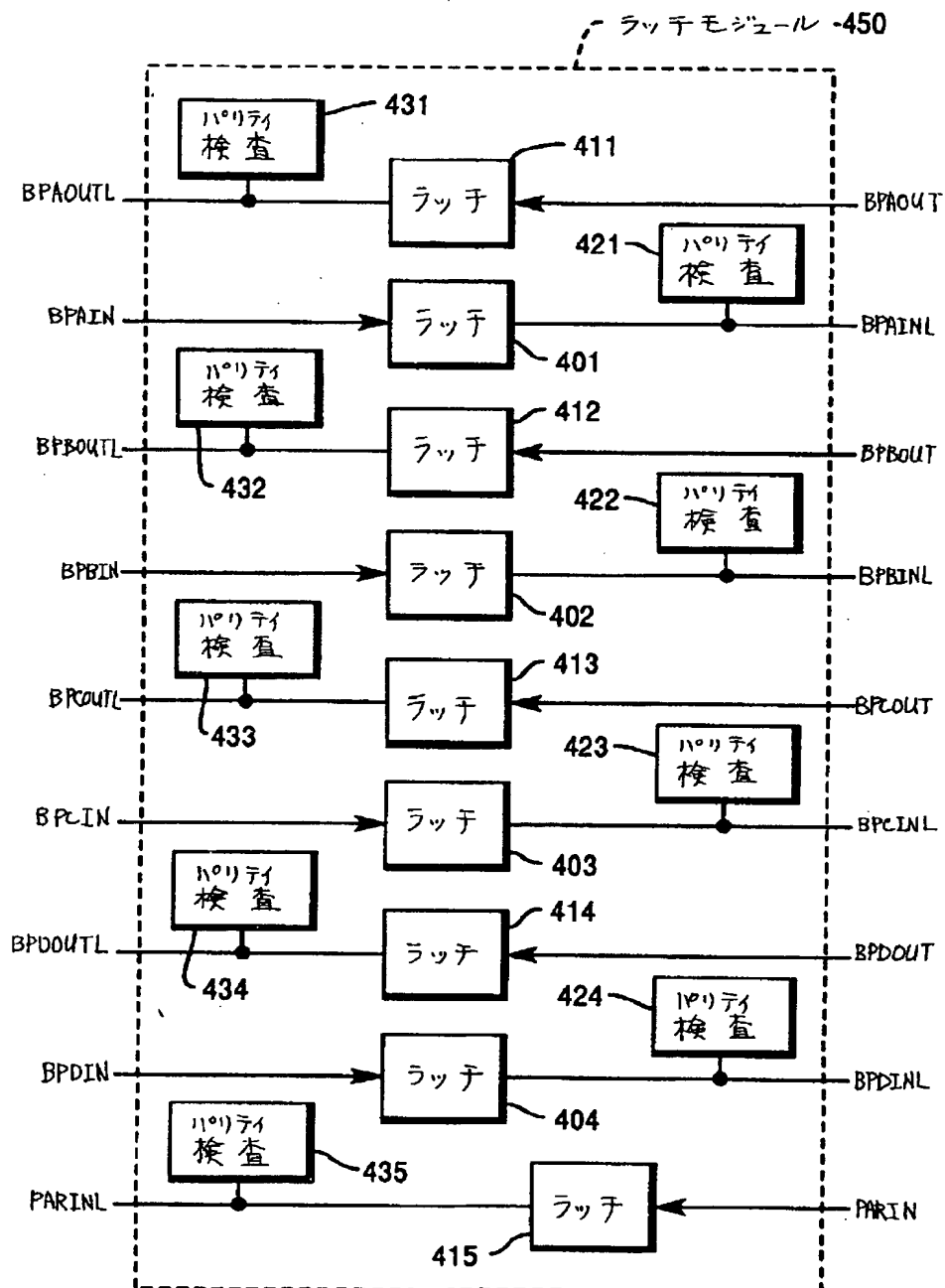
【図6】



(17)

特開平5-197495

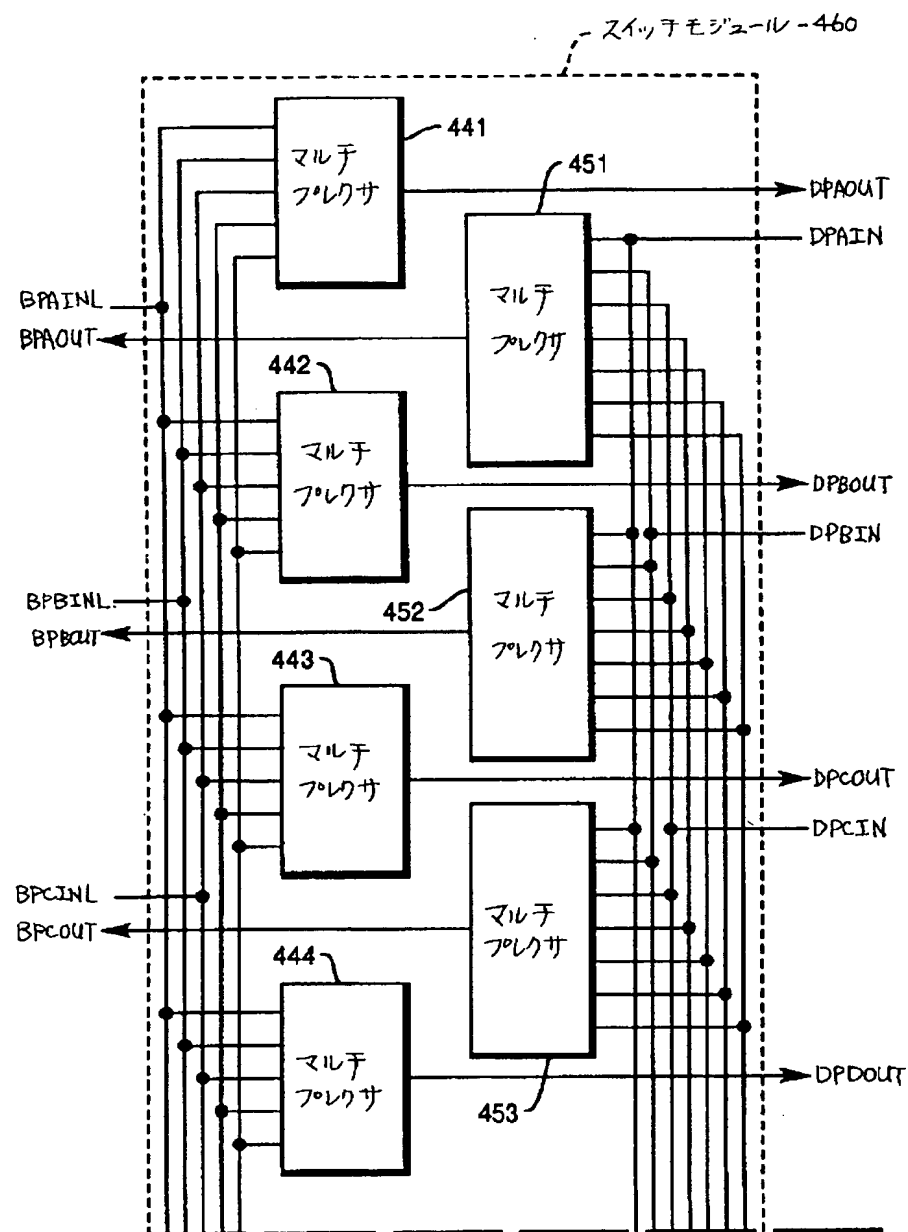
【図7】



(18)

特開平5-197495

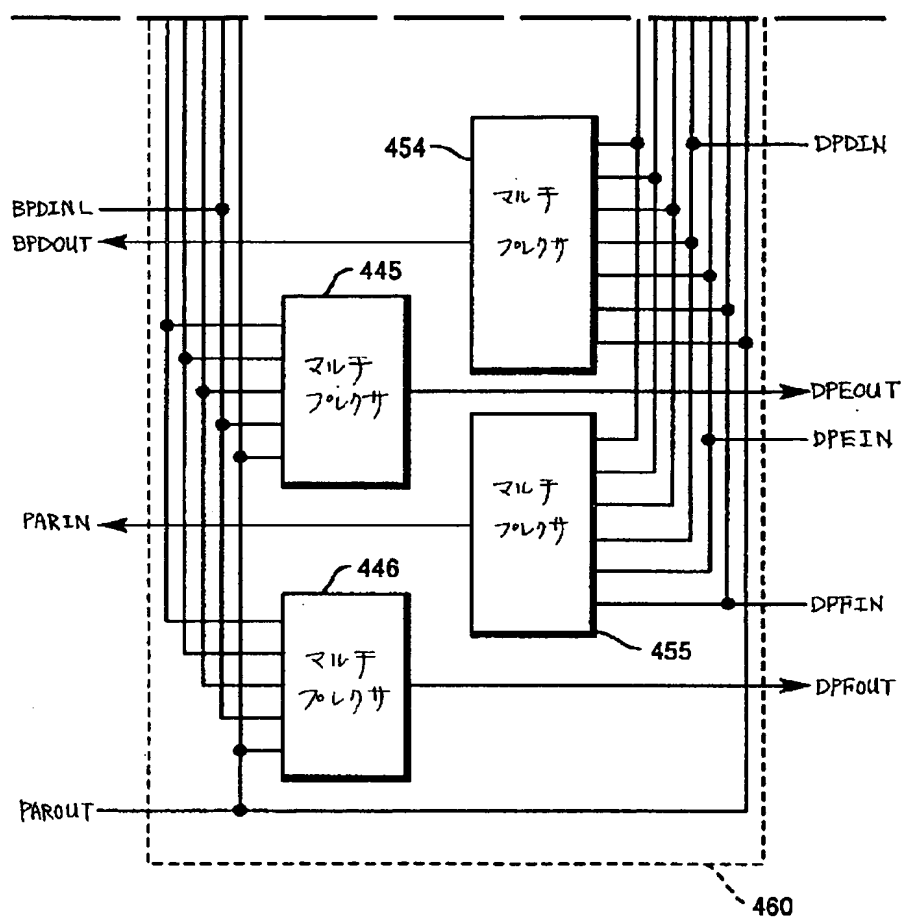
【図8】



(19)

特開平5-197495

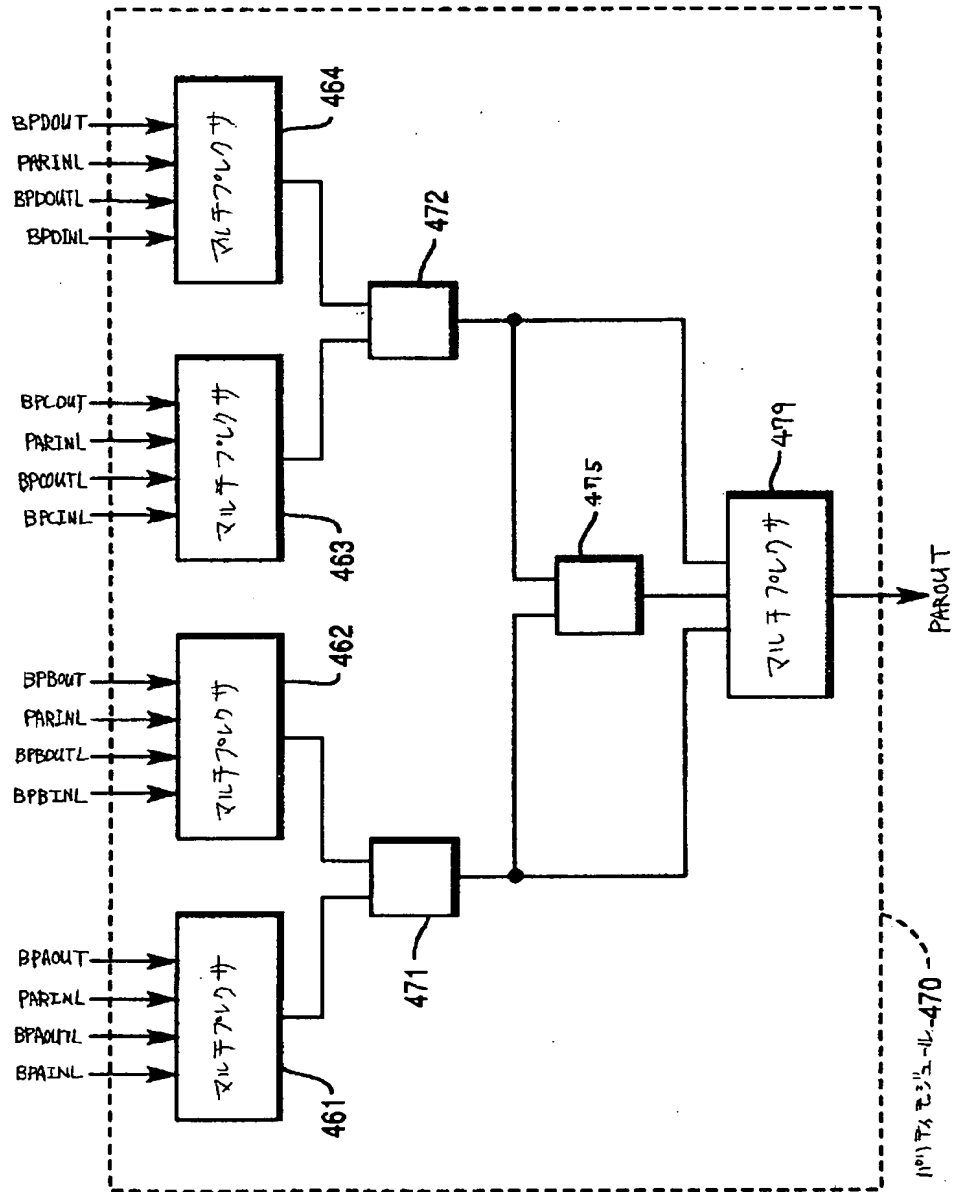
【図9】



(20)

特開平5-197495

【図10】



(21)

特開平5-197495

【図11】

